Building a National Microbial Pathogen Data Resource

Rick Stevens

The University of Chicago

Argonne National Laboratory

Who is NMPDR?

Lead by Rick Stevens (UC) and Ross Overbeek (FIG)

- The University of Chicago and Argonne
  - Terry Disz, Bob Olson, Mike Kubal, Liz Marland, Natalia Maltsev, Ed Frank, Wen- Hsuing Li, Richard Quigg, Jonathan Silverstein, Evgeni Selkov, Mark Hereld
- The Fellowship for Interpretation of Genomes
  - Veronika Vonstein, Rob Edwards, Sveta Gerdes, Andrei Osterman, Michael Fonstein, Gordon Pusch, Bruce Perrillo
- The University of Illinois at Urbana-Champaign
  - Scott Lathrop, Stephanie Mclean

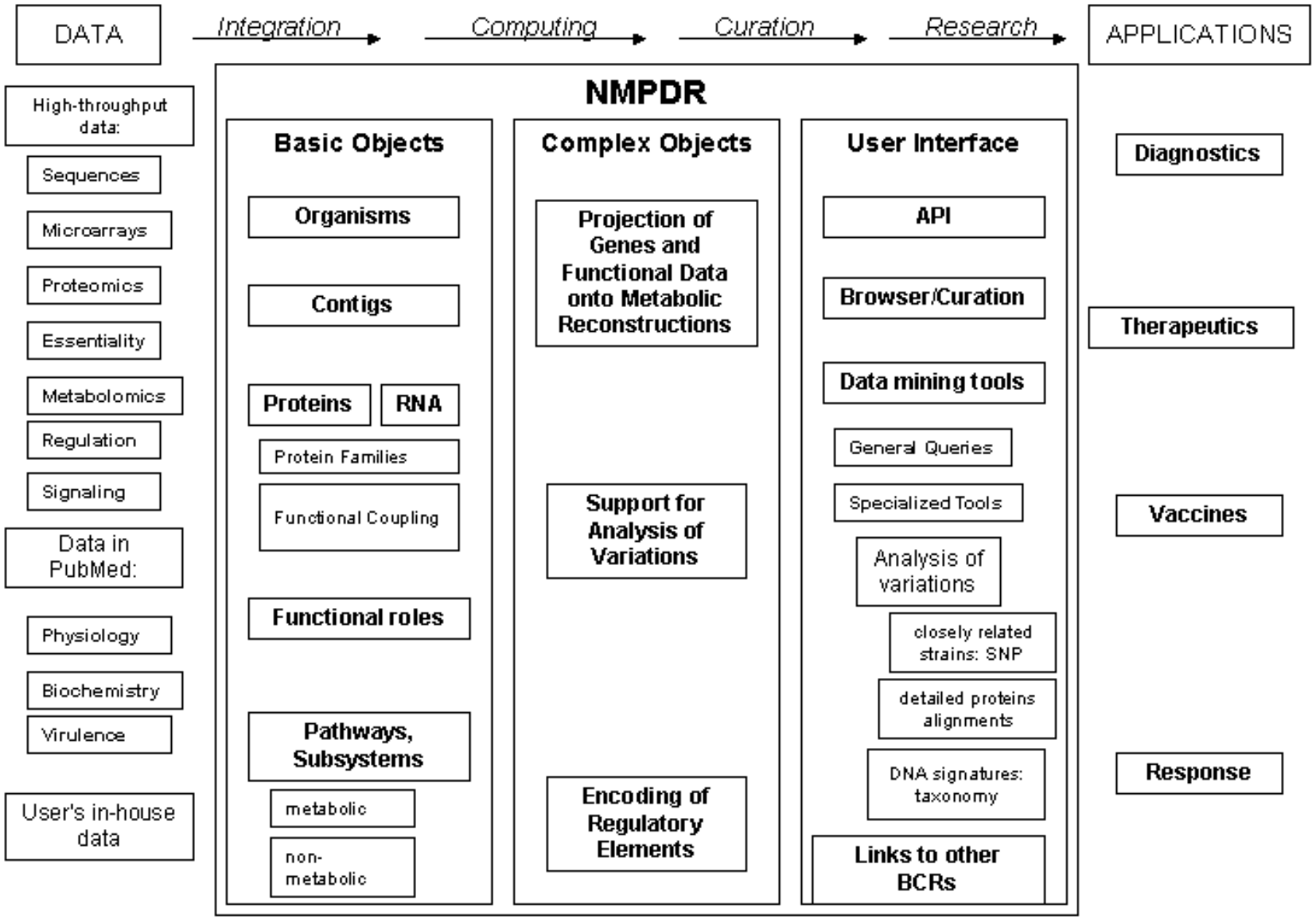
Our Value Proposition

- Our team has a strong history of integrated database development
  - GenoBase, WIT/PUMA, ERGO, SEED
- Strong focus on tools for comparative analysis
  - Evolutionary perspective
  - Functional coupling and analysis of conserved clusters
  - Horizontal annotation (subsystems)
  - Metabolic reconstruction
- Rapid DB development (ER modeling)
- Strong history of open source development and software packaging and distribution
- Coupling of computer science, bioinformatics and micro/molecular biology ⇒systems biology
- Large-scale and Grid computing - access to resources
- Collaboration and visualization technology

Our Vision

- A web based (browser and web services interfaces) gateway to all relevant data resources for supporting research in microbial pathogenesis

- The resource will be grounded on a well annotated set of genome sequences but will also include integrations with related datasets {protein structures, other gene products, proteomes, transcriptomes, metabolomes, derived data and models, etc.} including linkages to the primary literature



www.nmpdr.org

NMPDR National Microbial Pathogen Data Resource Center

Target Organisms

- *Staphylococcus aureus* 43546
- *Streptococcus pneumonia* 4637
- *Streptococcus pyogenes* 845
- *Vibrio cholera* 6274
- *Vibrio vulnificus* 622
- *Vibrio parahaemolyticus* 1354
- *Campylobacter jejuni* 3553
- *Listeria monocytogenes* 8067
- 23 genome sequences currently available

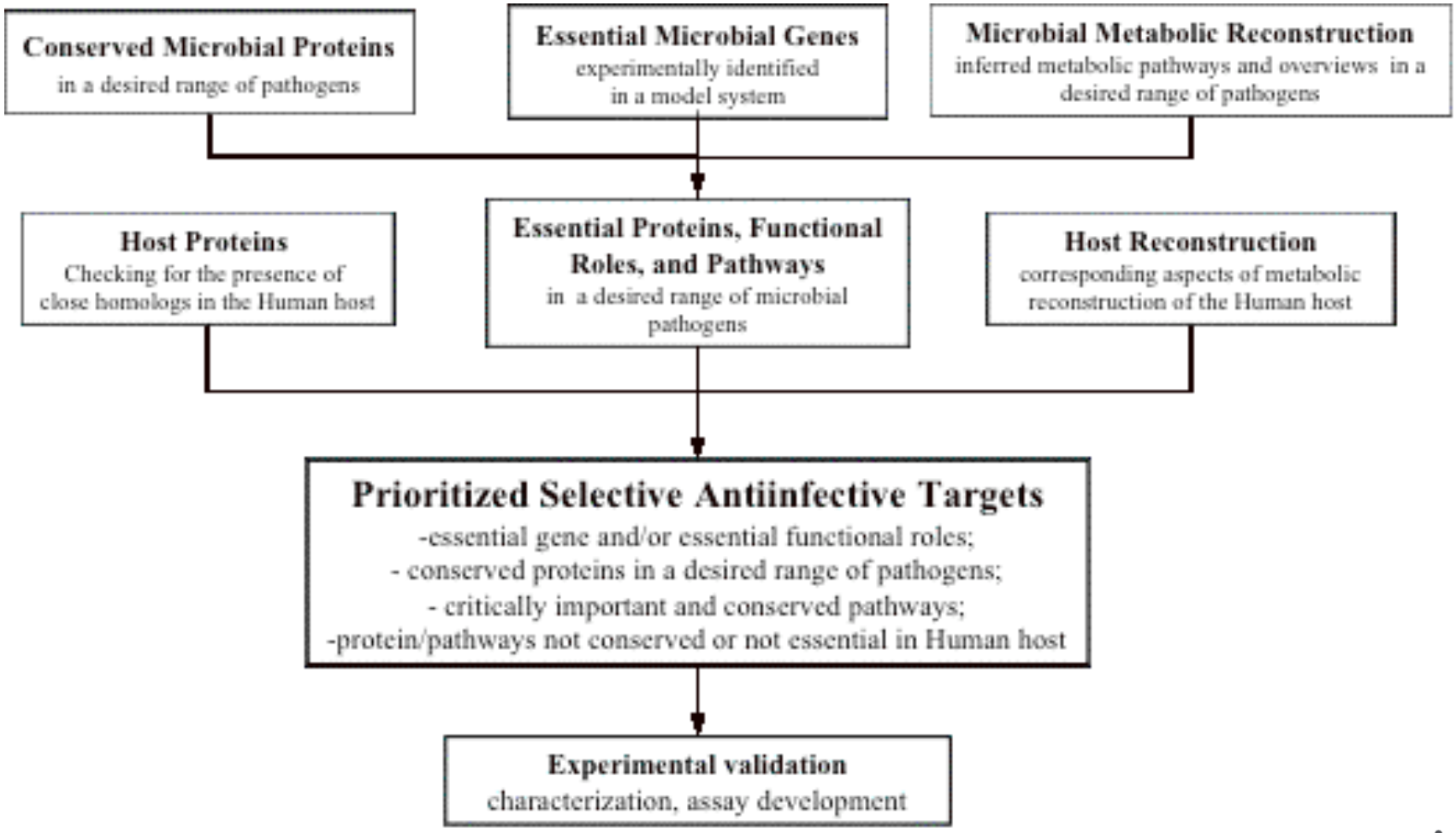
Organization Strategy

- Five major areas of activity
  - Production db and user support 3 FTE
  - Database and informatics tool development 5 FTE
  - Annotation and curation teams 5 FTE
  - Data acquisition and integration 2 FTE
  - Education, Outreach and Training 2 FTE
- Scientific Working Group
  - Organism related expertise, strategic directions
- User “advisory” Committee
  - Requirements generation, feedback, tactical priorities

Driving Applications

- We believe that system development should be balanced between technology push (what is possible) and applications pull (what is needed)
- Example drivers
  - Fundamental microbiology research (e.g. Dusko Ehrlich)
  - Anti-infective research (e.g. Olaf Schneewind)
  - Variation identification and strain based approaches to understanding antibiotic resistance
  - Host-pathogen interactions and immunology

Informatics Driven Approach to Anti-microbial Agent Development



www.nmpdr.org

NMPDR
National Microbial
Pathogen Data
Resource Center

Support for Studying Variation

- Tools for data exploration and comparative analysis of many (100's) of closely related strains
  - In the context of other pathogens and non-pathogens
  - In evolutionary (phylogenetic) context
  - With tools for the analysis of gene histories and horizontal gene transfer

Example Use Cases

- Comparative analysis of gene clusters
- Searching for missing genes
- Comparing pathways between organisms
- Projecting essentiality data between organisms
- Reconstructing virulence pathways
- Extracting rules for model development
- Visualization and browsing of pathways and networks
- Studying horizontal gene transfer of pathogenticity factors
- Studying evolution of antibiotic resistance

Conceptual Architecture:
A Layered System That Provides

- User level view of multiple entry points
  - Google like search across the elements of the resource
  - Organism (strain) specific databases
  - Organism cluster (related strains) browser
- Lightweight db server (data mining repo)
  - Mostly read only, easily extensible, integrated comparative analysis server and cluster browser
  - DAS compliant
- Organism specific db and pathway browser
  - Framework for navigating reconstructions of individual organisms, pathway based navigation
- General purpose p2p annotation system
  - Read/write expert curation interface, rapid prototyping environment, data mining, medium performance

Connection Opportunities

- Creation of a true network of BRCs
- Rapid response network for national crisis
- Cooperation in support of user communities
- Sharing and leveraging of tools and local expertise
- Identification of leveraging points
- Strategies for user interaction with BRCs

User Engagement Models

- Professional events, live tutorials, on-line, web and grid based training
- Visitor program, post-doc rotations, practicum etc.
- BRC x BRC comparisons and feature/function pressure
- Cross center challenge problems
- How much of these can/should be BRC wide?

Some Challenges We Face

1. Coupling to reasonable driver problems
2. Inter-operability and inter-ontologies
3. Exploiting web services effectively
4. Rational approach to sharing/layering and leveraging while preserving the ability for all to innovate
5. Presenting a unified capability to the user community

Some Modest Proposals

- Prokaryote Ontologies (starting with GO)
- Data Exchange (starting with annotations)
- Remote Access to Databases (starting with DAS)
- BRC Visitor Program (joint with RCEs?)
- ASM, SGM etc. collective outreach
- Peer Reviewed Subsystems e-Publications
- Library of Virulence/Pathogenicity Factors
- BRC Rapid Response Network